MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

A.I. Memo No. 510

February, 1979

DIFFERENTIAL GEOMETRY, SURFACE PATCHES AND CONVERGENCE METHODS

W.E.L. Grimson

ABSTRACT. The problem of constructing a surface from the information provided by the Marr-Poggio theory of human stereo vision is investigated. It is argued that not only does this theory provide explicit boundary conditions at certain points in the image, but that the imaging process also provides implicit conditions on all other points in the image. This argument is used to derive conditions on possible algorithms for computing the surface. Additional constraining principles are applied to the problem; specifically that the process be performable by a local-support parallel network. Some mathematical tools, differential geometry, Coons surface patches and iterative methods of convergence, relevant to the problem of constructing the surface are outlined. Specific methods for actually computing the surface are examined.

Acknowledgements: This report describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the laboratory's artificial intelligence research is provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research contract N00014-75-C-0643 and in part by National Science Foundation Grant MCS77-07569.

1. Introduction

In a recent article, Marr & Poggio [1977] set out a computational theory of human stereo vision. One consequence of this theory is that the stereo algorithm can at best determine disparity only along certain contours in the image. Two important questions thus become: Is it possible to reconstruct a surface which is consistent with the stereo information, and if so, how does one reconstruct such a surface? This paper addresses these questions and discusses several pieces of mathematics relevant to the interpolation of surfaces under such conditions.

The original motivation for the problem lies in the theory of human stereo vision; thus, we seek an algorithm which is plausible as a human model, although we will not claim that the resulting algorithm is an exact model of a module of the human system.

In designing a plausible algorithm, a number of constraining principles are applied to the problem, and are accepted as given. Any visual system must be able to process large amounts of input data; for example, in the human system, the central part of the visual field may be considered as an image which is a thousand pixels on a side. [1] At the same time, it is desired that the system perform its computations in real time; i.e. that it take only a short period of time after receiving the input to complete its computation. Since each processor must take some non-negligible amount of time to perform each action, this essentially implies the use of computations which can be implemented in a parallel manner, using a large number of interconnected processors. This is the first constraint.

A second physical constraint on our process is that the "hardware" which implements it must fit into a finite and constrained amount of space (such as the cranium). The use of parallel networks of interconnected processors, combined with this constraint of physical space available, requires that each processor not be connected to all others. Rather, there should only be local connections between the processors. Here, local means not only that the number of connections be small, but that since we are processing information whose underlying coordinate system is a two-dimensional plane, the connections should also be local in a spatial sense. If the support of a function, defined on a two-dimensional grid, is the set of points on the grid which contribute in a non-trivial manner to the computation of the function,

---

1. The size of the individual receptors in the center of the foveal region is roughly 20 to 35 seconds of arc (T.N. Cornsweet, Visual Perception [1970], p. 356). Over the central 6 degrees of the visual field, this implies a figure on the order of roughly 1000 receptors.

then our requirement is that the processors implementing our computation must have local (topological or metric) support.

A third requirement, though not as strong as the first two, concerns the complexity of the individual processors. We do not want to achieve locality at the expense of computation time. That is, if the computation can be made local only at the expense of requiring each processor to compute some complex function requiring a long time interval to complete, then there is something fundamentally wrong with the computation as a part of the image processing system. Note that the global computation performed by the parallel network need not be simple, it is only necessary that the individual processors not be complex. Connected with this desire for simplicity is a second desire for uniformity, that is, if possible, the individual processors in the network should be identical. However, this is not as critical a requirement as that of parallelism and local support.

Although the original motivation for such constraints on the algorithm arise from consideration of the human visual system, they could apply equally well to other types of image processing systems, and are taken as general constraints on the computation we are about to investigate. As a consequence, this paper will not suggest a particular algorithm as a model of the human system. Rather, a number of alternatives will be suggested and examined in terms of their acceptability vis-a-vis the algorithmic constraints outlined above.

The problem is approached in the following manner. We first determine exactly what information is available from the stereo algorithm. It is shown in the second section that the stereo algorithm gives implicit as well as explicit information about the shape of the surfaces in a scene. Next, we turn this information into conditions and constraints on the computation. In section three we outline general methods for satisfying these constraints. This includes an outline of some differential geometry relevant to the description and reconstruction of surfaces. Also, several extremal conditions for the interpolation of a surface are suggested. Section four outlines the Coons surface patch method for creating smooth fair surfaces from boundary conditions. Finally, we design an actual algorithm which satisfies the algorithmic constraints of simple, parallel, local-support processes and which solves the problem subject to the computational constraints developed in the previous sections. Section five outlines methods for performing the computation. Section six combines the previous sections and illustrates how to actually construct the surface.

2. Information from Stereo

According to the current computational theory of human stereo vision (Marr & Poggio [1977]), the human visual processor solves the stereoscopic matching problem by means of an algorithm that consists of five main steps: (1) The left and right images are each filtered at different orientations with bar masks of four sizes that increase with eccentricity; these masks have a cross-section that is approximately the difference of two gaussian functions, with space constants in the ratio 1:1.75. (2) Zero-crossings in the filtered images are found, along scan-lines lying perpendicular to the orientation of the mask. Termination points of lines and edges are also localized. (3) For each mask size, matching takes place between pieces of zero-crossing contour of the same sign and roughly the same orientation in the two images, for a range of disparities up to about the width of the mask's central region. Within this disparity range, Marr & Poggio showed that false targets pose only a simple problem. (4) The output of the wide masks can control vergence movements, thus causing small masks to come into correspondence. In this way, the matching process gradually moves from dealing with large disparities at low resolution to dealing with small disparities at high resolution. (5) When a correspondence is achieved, it is stored in a dynamic buffer, called the 2 1/2-dimensional sketch.

The justification for this model of stereo processing is well detailed in Marr & Poggio [1977], and will not be dealt with here. Our concern in this paper is how to use the information provided by the stereo algorithm to construct a representation of the underlying surfaces in the image. To do this, one must carefully consider what information is actually provided by the stereo mechanism.

The important features of the stereo theory, from the point of view of this paper, are the following. Each image is convolved with a mask which is a two-dimensional difference of gaussians.[1] The convolutions are searched for zero-crossings and the position and sign of such zero-crossings are stored in a description of the image. These zero-crossing contours form the basic descriptors which are matched by the stereo algorithm. As a consequence the only places in the image which may have explicit disparity information associated with them are those corresponding to zero-crossing contours.

---

1. In the actual implementation of the stereo theory, a circular symmetric difference of gaussians mask is used rather than an oriented mask. The justification for the use of such masks is found in Marr & Hildreth [1979].

There is strong psychophysical evidence for such operators (Marr & Poggio [1977], Wilson & Giese [1977], Wilson & Bergen [1979]). What is the computational motivation for such operators? In the one-dimensional case, such an operator detects an intensity change; specifically, the points of inflection in intensity. For the two-dimensional case, we again want to detect those points which correspond to a point of inflection, now for some directional derivative. Marr & Hildreth [1979] have shown that the desired orientation of the directional derivative should be such as to coincide with the local orientation of the underlying line of zero-crossings. Under certain conditions, this orientation is the one at which the zero-crossing has a maximum slope and this can be detected using an orientation-independent differential operator, the Laplacian ∇^2. Thus, these operators[1] essentially detect inflections in intensity for some resolution - the smaller the mask's central excitatory region, the sharper the inflection must be in order to be detected. Note that these are not discontinuities in an analytic sense, but rather these are "discontinuities" with respect to some resolution of the image.

From what do such inflections in intensity arise? Image formation is affected by four factors (Woodham [1978]):

- imaging geometry
- incident illumination
- surface photometry
- surface topography.

Surface photometry refers to how light is reflected by the object surface. It is determined by optical constants of the object material and by the surface microstructure (detail too fine to be resolved but which causes observable effects in the way light is reflected at the object surface). Surface topography is the surface detail which is within the resolution limits of the imaging hardware. It refers to the gross object shape relative to the viewer. If we assume some illumination and imaging geometry, then inflections in the image intensities will be caused either by the surface photometry or the surface topography.

Thus sharp changes in color or reflectance of the surface, scratches in the surface, sharp changes in the shape of the surface all can give rise to intensity inflections. These intensity inflections will cause zero-crossings in the convolutions and it is these primitive descriptors which are matched by the stereo algorithm. Given the fact that inflections are caused by such effects of the surface photometry

---

1. The details of such "edge detectors" may be found in Marr & Hildreth [1979].

and topography, one consequence of the stereo algorithm is that at best it returns disparity values along some set of contours in the image. One may explicitly determine depth or surface orientation only along such contours. Our task is to reconstruct a description of the surface (either in depth or in surface orientation) at all points in the image.

In general, any one of a multitude of widely varying surfaces could fit the boundary conditions imposed by the stereo algorithm. But to be completely consistent with the stereo process, such surfaces must meet the depth or surface orientation conditions along the zero-crossing contours and not give rise to any other zero-crossing contours which do not appear in the convolved image. This is captured by the following assertion:

Assertion: **Places of no information are actually places of information.**

By this, we mean that for the locations of the image not associated with a zero-crossing, we may assume that the underlying surface does not change in a radical way. An intuitive way of looking at this claim is as follows. Suppose we are given a closed zero-crossing contour, within which there are no other zero-crossings. An example would be a circular contour, along which the disparity is constant. One surface which is consistent with this set of boundary conditions is a sphere. However, one could also fit a highly convoluted surface (for example sin(r)/r) to this set of boundary values. Yet in principle, such a rapidly varying surface should give rise to other zero-crossings, since the intensity across the surface should also vary considerably. Hence, we claim that the set of zero-crossing contours contains implicit information about the surface as well as explicit information, and such information can be valuable in reconstructing the surface.

This assertion is equivalent to the following statement:

Assertion: **Except under certain singular conditions, a rapid change in the direction of curvature of a surface must give rise to an inflection in intensities, i.e. you cannot hide a dip.**

In order to prove this assertion, we shall need to first develop tools for dealing with image formation. This will allow us to relate changes in the surface orientation of a surface element to changes in the perceived intensities. Then we may state the problem formally as a set of lemmas dealing with one and two dimensional cases.

Since the intensities with which we shall deal are caused by both surface photometry and surface topography, it is possible to have complex interactions between the two effects. In the case in which both factors have roughly equivalent effects, one could construct situations in which the surface topography changes radically, yet the surface photometry also changes sufficiently so that there are no noticable changes in the image intensities. In such situations, it seems unlikely that one can determine any information about the shape of the surface from the image intensities themselves. We concentrate instead on those situations in which the changes in the surface topography dominate changes in the surface photometry.

In order to prove this assertion, we develop some tools for dealing with image formation [Horn 1970, 1975]. In the simplest case of a single point source, the geometry of reflection is governed by three angles. The incident angle between the local normal of a surface portion and the incident ray is called i, the view angle between the local normal and the emitted ray is called e, and the phase angle between the incident and emitted rays is called g. The fraction of incident illumination at a given surface point reflected in the direction of the viewer is denoted by the reflectance function $\phi(i,e,g)$. Most situations with more complicated distributions of light sources can be modelled by the superposition of single point sources.[1]

We let $\Lambda(x,y,z)$ be the object irradiance at the surface point (x,y,z), scaled by the ratio of image irradiance to scene radiance. The object irradiance will be constant or obey some inverse-square law with respect to distance from the source for physical systems. The ratio of image irradiance to scene radiance is a constant which depends on the imaging system.

Let $\mathbf{r} = (x,y,z)$ be a visible point on an object, and $\mathbf{r}'$ be the corresponding point in the image. If $b(\mathbf{r}')$ is the image irradiance measured at the image point $\mathbf{r}'$, then (Horn [1970]):

$$b(\mathbf{r}') = A(\mathbf{r})\phi(i,e,g).$$

If we restrict our attention to situations in which the light source can be considered distant relative to the separation of object and viewer, then the phase angle is roughly constant and $\phi(i,e,g)$ can be replace by the radiance function $R(p,q)$, where $p = z_x$ and $q = z_y$ are the partial derivatives of the surface with respect to the two coordinate variables, x and y. It may be that we can in many cases decompose

---

1. For a more complete development of the mathematics of image formation, see Horn [1970, 1975], Woodham [1978].

the surface photometry from the surface topography in such a manner, so that the image intensities depend only on p and q. In fact, Woodham [1978] observes:

> " No matter how complex the distribution of incident illumination, for most surfaces, the fraction of the incident light reflected in a particular direction depends only on the surface orientation."

If we also restrict our attention to situations in which the image projection is orthographic, then

$$b(r')=I(x,y)$$

where I(x,y) is the intensity value recorded in the image. Thus, the image equation is given by:

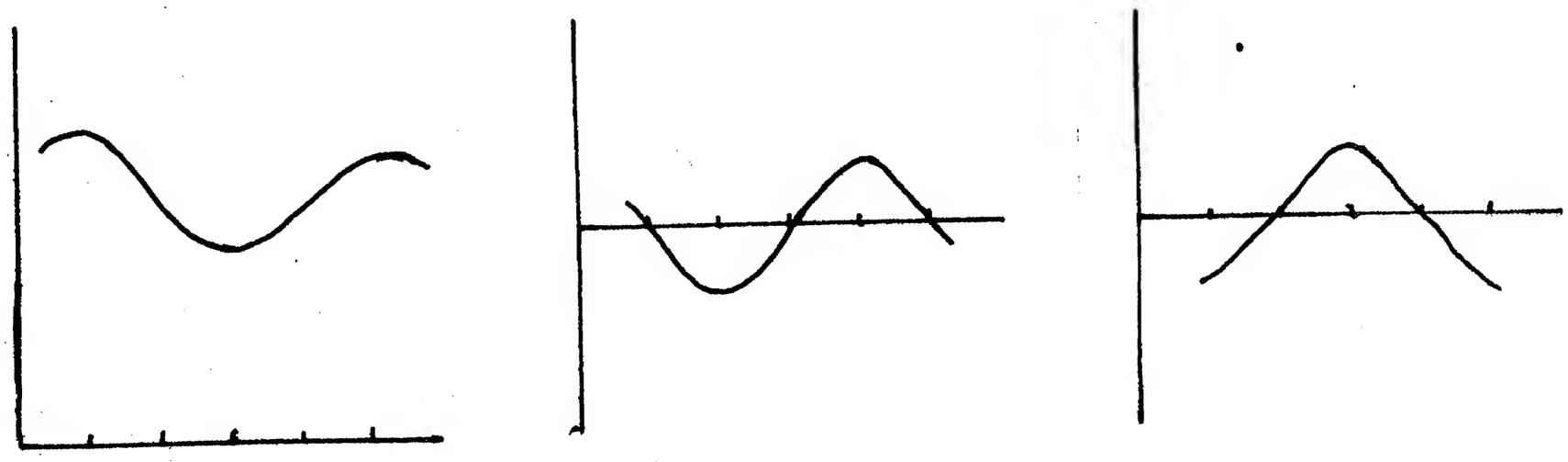
$$I(x,y) = A(x,y,z)R(p,q).$$

Note that in the case of uniform illumination and uniform photometric properties of the surface, A is constant. However, in general, the intensities will be a function of the position as well as the orientation of a surface patch.

To prove the assertion, we first examine the one dimensional case. We wish to investigate the conditions under which a bending of the surface forces an inflection in the intensities. In what follows, we assume that A, R and z are second differentiable functions. We begin with the following lemma:

Lemma: Suppose that the surface and its reflective properties are such that changes in the orientation of the surface dominate changes in intensity due to changes in position on the surface. If the surface contains at least two inflection points, then unless the reflectance R is constant for the values of p involved, the intensities must contain an inflection.

Proof: Consider a second differentiable surface z(x) which contains at least two adjacent inflection points, at x_1 and x_2. Then z_{xx} and consequently p_x are zero at these points.



Differentiating the image intensity equation yields

$$I_x(x) = A_x(x)R(p(x)) + A(x)R_p(p(x))p_x(x).$$

Note that if the reflectivity is completely independent of the position in the image, then A is constant and hence $A_x = 0$. Then, since $p_x(x_1) = p_x(x_2) = 0$, we see that $I_x(x_1) = I_x(x_2) = 0$. Since $I_x(x)$ has two adjacent zero points, the second differentiability of the surface z and the functions R and A thus implies that I_x must have an extremum for some value of x_0, $x_1 < x_0 < x_2$. In other words, I(x) must have an inflection at this point. This is provided $I_x(x)$ is not zero everywhere, and this is true provided that R_p is not zero everywhere, or that R is not constant.

For the more general case of the reflectivity being a function of the position x, as well as a function of the orientation p, we assume that

$$|A_x/A| << |R_p p_x| \text{ almost everywhere}$$

In other words, the major cause of change in intensity is changes in orientation, so that the surface shape changes much faster than the reflectivity.

Consider some neighbourhood of the point x_1. Because $p(x_1)$ is an extremum, there are constants α_1, $\alpha_2 > 0$, such that $p(x_1-\alpha_1) = p(x_1+\alpha_2)$ and such that

$$\text{sgn}\{I_x(x_1-\alpha_1)\} = \text{sgn}\{A(x_1-\alpha_1)R_p(p(x_1-\alpha_1))p_x(x_1-\alpha_1)\}$$
$$\text{sgn}\{I_x(x_1+\alpha_2)\} = \text{sgn}\{A(x_1+\alpha_2)R_p(p(x_1+\alpha_2))p_x(x_1+\alpha_2)\}$$

Here, sgn(x) is 1 if x is positive, -1 if x is negative and 0 if x is zero.

Using the fact that

$$\text{sgn}(xy) = \text{sgn}(x)\text{sgn}(y),$$
$$p(x_1-\alpha_1) = p(x_1+\alpha_2),$$

and that A(x) is non-negative, we see that

$$\text{sgn}(I_x(x_1-\alpha_1)) = \text{sgn}(I_x(x_1+\alpha_2))$$

if and only if

$$\text{sgn}(p_x(x_1-\alpha_1)) = \text{sgn}(p_x(x_1+\alpha_2)).$$

But since x_1 is an inflection point, this implies that

$$\text{sgn}(p_x(x_1-\alpha_1)) \neq \text{sgn}(p_x(x_1+\alpha_2)).$$

By the second differentiability of I, I_x must be zero for some point in this neighbourhood.

Similarly, we may show that I_x is zero for some point in a neighbourhood about x_2. Arguing as before, we may then show that I(x) contains an inflection point except in the case of R being constant.

Lemma: Suppose that the surface and its reflective properties are such that changes in the orientation of the surface dominate changes in intensity due to changes in position on the surface. If the surface contains one inflection point at x_1 and R(p) has an extremum for some value p_0 such that $p(x_2) = p_0$, where $x_2 \neq x_1$, then unless the reflectance R is constant for the values of p involved, the intensities must contain an inflection.

Proof: The proof of this lemma is very similar to the previous one. As before, we may show that $I_x = 0$ for some point in a neighbourhood about x_1. Consider a neighbourhood of the point x_2. As before, we may find constants α_1 and α_2 such that

$$\mathrm{sgn}\{I_x(x_2-\alpha_1)\} = \mathrm{sgn}\{A(x_2-\alpha_1)R_p(p(x_2-\alpha_1))p_x(x_2-\alpha_1)\}$$
$$\mathrm{sgn}\{I_x(x_2+\alpha_2)\} = \mathrm{sgn}\{A(x_2+\alpha_2)R_p(p(x_2+\alpha_2))p_x(x_2+\alpha_2)\}.$$

In this case, since there is no inflection in the surface here,

$$\mathrm{sgn}(I_x(x_2-\alpha_1)) = \mathrm{sgn}(I_x(x_2+\alpha_2))$$

if and only if

$$\mathrm{sgn}(R_p(p(x_2-\alpha_1))) = \mathrm{sgn}(R_p(p(x_2+\alpha_2))).$$

But since R has an extremum at $p(x_2)$,

$$\mathrm{sgn}(R_p(p(x_2-\alpha_1))) \neq \mathrm{sgn}(R_p(p(x_2+\alpha_2))).$$

Arguing as before, we see that I(x) must therefore contain an inflection point.

The only other possible case of the surface containing an inflection point is if R is monotonic and the surface has only one inflection point. In thius case, it is not possible to guarantee that there be an inflection in the intensities.

Stated in less formal terms, the above lemmas show that if the major changes in the intensities associated with a particular surface are due to the shape of the surface and not due to changes in the photometry of the surface, then certain types of changes in the surface shape must give rise to zero-crossings. This means that even if the surface is painted with a varying shade of color, so long as the effect of the shading does not overwhelm the actual bending of the surface, there must be an inflection in the intensities corresponding to the change in shape of the surface. Of course, it may be possible to paint a curved surface in such a way as to exactly counteract the effects of the curving of the surface on the intensities, thereby ensuring that no zero-crossing will correspond to the change in the surface shape. We regard such situations as special cases which will be rare in the real world, since they require a peculiar coupling of the surface topography and photometry which is dependent on the viewer and illumination geometry.

The lemmas can be extended to the two-dimensional case. The proofs are outlined below.

Lemma: Suppose that the surface and its reflective properties are such that changes in the orientation of the surface dominate changes in intensity due to changes in position on the surface. Furthermore, suppose that along some direction, the surface undergoes radical changes of shape, while perpendicular to this direction, the change in surface shape is negligible. If the surface contains two inflection points, then unless the reflectance R is constant for the values of p involved, the intensities must contain an inflection.

Proof: The proof is similar to that of the one-dimensional case. Without loss of generality, we assume that the surface inflections occur along a line parallel to the x axis. In this case,

$$I_x(x,y) = A_x(x,y)R(p,q) + A(x,y)R_p(p,q)p_x(x,y) + A(x,y)R_q(p,q)q_x(x,y).$$

Since the direction of concern is parallel to the x axis, y is constant along such a slice of the surface, say $y = y_0$. As before, we wish to show that $I_x(x,y_0)$ is zero for some point in a neighbourhood of the inflection point x_1 and for some point in a neighbourhood of the inflection point x_2. Since we have assumed that the effects of p_x dominate those of q_x, we can essentially use the same argument to show that I_x changes sign within the neighbourhood and hence must be zero at some point within it.

Lemma: Suppose that the surface and its reflective properties are such that changes in the orientation of the surface dominate changes in intensity due to changes in position on the surface. Furthermore, suppose that along some direction, the surface undergoes radical changes of shape, while perpendicular to this direction, the change in surface shape is negligible. If the surface contains one inflection point at x_1 and R(p) has an extremum for some value p_0 such that $p(x_2) = p_0$, where $x_2 \neq x_1$, then unless the reflectance R is constant for the values of p involved, the intensities must contain an inflection.

The proof again parallels that of the one-dimensional case, with modifications similar to those used in the proof of the above lemma.

We have shown that except under certain singular conditions, a pair of inflections in the surface must cause an inflection in intensities. So at some level of resolution, the convolutions of the image will detect such inflections (if they are large enough) as zero-crossings. Hence the stereo output tells us not only the shape

of the surface along the zero-crossings, but also that the surface cannot curve rapidly or drastically between zero-crossings. Otherwise, an inflection would cause a zero-crossing and none are evident.

Thus, we claim that in general, places which do not have a zero-crossing contour must be "well-behaved" in the sense of curving as little as possible and of having the sign of the curvature change as little as possible. In the next section, we make precise the notion of curvature. This will allow us to use our assumption about places without zero-crossings to design algorithms for reconstructing a surface.

3. Differential Geometry

In order to develop methods for reconstructing the surface, we need to relate the shape of the surface to inherent properties of the surface. More importantly, we need to relate the shape of the surface to conditions on the surface orientation at each point of the surface or to conditions on the depth values at each point of the surface. To do this, we first review some relevant pieces of differential geometry. For the most part, these are taken from Weatherburn [1927].

We have already indicated that the surface (or piece thereof) with which we are dealing has no discontinuities in depth and is smooth. This suggests that the reconstruction of the piece of surface is intimately related to the curvature of the surface. We now make the notion of the curvature of a surface more precise.

3.1 Curvature of Curves

For a curve, the curvature at a point, κ, is defined in the following manner.

A curve is the locus of a point whose position vector **r** (relative to a fixed origin) is a function of one parameter. In particular, that parameter can be taken to be the arc length of the curve, s, measured from a fixed point on the curve.

The (unit) tangent to the curve at a point is defined by

$$\mathbf{t} = d\mathbf{r}/ds .$$

In the case where the parameter of the curve is taken to be arc length, the tangent vector given above is automatically a unit vector.

The curvature at any point on the curve is then defined as the arc-rate of rotation of the tangent. It is also given more formally by the relation

$$d\mathbf{t}/ds = \kappa \mathbf{n}$$

where **n** is a unit vector perpendicular to **t** and lies in the plane spanned by the tangents at that point and a consecutive point. In other words, let P be the point associated with the arc length value s, and let P_1 be the point associated with the arc length value s+ds. As ds tends to zero, the tangents at the points P and P_1 will define a plane, such that both tangents lie in this plane. The vector **n** lies in such a plane, and is perpendicular to **t**. This vector, **n**, is the principal normal of the curve at that point, and the plane containing two consecutive tangents at P, is called the osculating plane.

An alternative method for defining the curvature of a curve at a point is as follows. The circle of curvature at a point P on the curve is the circle passing through three points on the curve which in the limit coincide at the point P. In other words, consider a point P on the curve specified by some value of the arc length, say s_0. Let P_1 be the point on the curve associated with the arc length value $s_0+\epsilon$ and let P_2 be the point associated with the arc length value $s_0-\epsilon$. For any ϵ these three points define a circle passing through all three points. As we let ϵ tend to zero, the associated circle converges to the circle of curvature. Its radius ρ is related to the curvature of the curve at that point by

$$\rho = 1/\kappa .$$

Expressed in intuitive terms, the curvature at a point measures how quickly the curve deviates from a straight line.

The particular normal to the curve at P that is perpendicular to the osculating plane is the binormal. Being perpendicular to the osculating plane means that it is perpendicular to both t and n, and hence is parallel to txn. This unit vector is denoted b.

Just as the curvature κ measures the arc rate of turning for the unit vector t, the torsion τ measures the arc rate of turning for the unit vector b. It can be obtained from the relation

$$db/ds = -\tau n .$$

In intuitive terms, the torsion measures how quickly the curve deviates from a plane.

3.2 Curvature of Surfaces

Now we can turn to surfaces and again define the notion of curvature.

A surface is the locus of a point whose coordinates are functions of two independent parameters, u and v. Thus the parametric equations for a surface, defined in a Cartesian coordinate system are

$$x=f_1(u,v) \quad y=f_2(u,v) \quad z=f_3(u,v)$$

and the surface is defined by some function $F(u,v)=0$.

Consider any curve drawn on the surface. Again, let s be the arc length of the curve, measured from some fixed point on the curve to the current point {x,y,z}. Then the tangent to the curve is the vector {x',y',z'}, where the ' refers to differentiation with respect to s. Now the straight line generated by the tangent to any curve is called a tangent line. In particular, all tangent lines at a point {x,y,z}

are perpendicular to the vector $\{F_x,F_y,F_z\}$. (Here we use the notation F_x to denote the partial derivative of F with respect to x.) To see this, note that F has the same value at all points of the surface, hence it remains constant along any curve as s varies. Thus

$$F_x\, dx/ds + F_y\, dy/ds + F_z\, dz/ds = 0 .$$

Thus $\{x',y',z'\}$ and $\{F_x,F_y,F_z\}$ are perpendicular. So all tangent lines at a point are perpendicular to this vector, and thus lie in a plane through $\{x,y,z\}$ perpendicular to this vector. This is the tangent plane. The normal to the plane at the point of contact is the normal to the surface at that point.

Any relation between the parameters, say f(u,v)=0 represents a curve on the surface since then r is a function of only one independent parameter. In particular, the parametric curves are those for which

$$u = \text{constant} \quad \text{or} \quad v = \text{constant}$$

Then if we denote

$$r_1 = \partial r/\partial u \qquad r_2 = \partial r/\partial v$$

we have that r_1 is a vector tangent to the curve v = constant at the point r. Similarly for r_2.

Consider two neighbouring points on the surface, with position vectors r and r+dr, corresponding to the parameter values (u,v) and (u+du,v+dv) respectively. Then

$$dr = r_1 du + r_2 dv .$$

Since the two points are arbitrarily closely spaced points on a curve passing through them, the length ds of the element of arc joining them is equal to the actual distance |dr| between them. Thus

$$ds^2 = r_1^2 du^2 + 2r_1 \cdot r_2 du dv + r_2^2 dv^2 .$$

We define

$$E = r_1^2$$

$$F = r_1 \cdot r_2$$

$$G = r_2^2 .$$

These quantities are called the fundamental magnitudes of the first order, and, together with the following quantity, are of use in computing characteristics of the surface.

$$H^2 = EG - F^2$$

By definition, the normal to the surface at any point is perpendicular to every tangent line through that point. Hence it is perpendicular to both r_1 and r_2.

Thus the unit normal is given by

$$n = r_1 x r_2 / H .$$

In a similar manner the second derivatives of r are denoted

$$r_{11} = \partial^2 r/\partial u^2 \quad r_{12} = \partial^2 r/\partial u \partial v \quad r_{22} = \partial^2 r/\partial v^2.$$

The second order magnitudes of the surface are then defined as

$$L = n \cdot r_{11}$$

$$M = n \cdot r_{12}$$

$$N = n \cdot r_{22}$$

$$T^2 = LN - M^2 .$$

We can now formalize the curvature of the surface itself. We have seen how one may define the curvature of a curve at a point. Consider any plane which intersects the surface at a particular point P, and which contains the normal to the surface at that point. The result of such a normal section is a curve, and we may evaluate the curvature of that curve at the point P. However, there are infinitely many planes through P which contain the normal to the surface at P. Can we identify any particular normal sections?

Given a point P on the surface, any curve on the surface passing through the point will have a tangent vector defined at the point. The plane containing all the tangent vectors for any curve passing through the point is called the tangent plane for that point. Suppose we intersect the tangent plane with the surface, and examine the rate at which the surface deviates from the plane along any particular direction. We will find that there are two directions on the surface, at right angles to each other, such that in one direction the surface deviates the quickest from the plane, and in the other direction the surface deviates the slowest. Both of these directions have the property that the normal at a consecutive point separated by an infinitesimal distance in either direction meets the normal at P. This means that the curve for a section along one of these directions has no torsion, and is subject to curvature in only one direction. These are the principal directions.

The values for the curvature of the curves obtained by taking normal sections along these principal directions are extrema. In other words as we change the direction of the section, the curvature for a normal section achieves a maximum at one of the principal directions. It achieves a minimum at the other principal direction. Let the curvatures of these special sections be κ_a and κ_b respectively. In the case of a plane, the curvature of each normal section is identical. In this case, any pair of perpendicular directions may be taken as the principal directions.

To relate the principal curvatures to the previous definition of curvature in terms of radii of circles, note the following. A curve on the surface such that the normals at consecutive points intersect is called a line of curvature. The point of intersection of consecutive normals along a line of curvature at P is the centre of curvature at P. Its distance from P is a principal radius of curvature and the reciprocal is a principal curvature.

Thus at each point there are two principal curvatures κ_a and κ_b. These are the normal curvatures of the surface in the directions of the lines of curvature. Given the principal curvatures, the curvature of the surface can be described in a number of ways. The first curvature of a surface is defined by

$$J = \kappa_a + \kappa_b.$$

The second curvature of a surface (also known as the specific curvature or the Gaussian curvature) is defined by

$$K = \kappa_a \kappa_b.$$

These are related to the fundamental magnitudes by

$$J = (EN-2FM+GL)/H^2$$

$$K = (LN-M^2)/H^2 .$$

In intuitive terms, the first curvature is analogous to the curvature of a curve, while the second curvature is analogous to the torsion of a curve.

Any point on the surface may be defined by the value of the Gaussian curvature at that point. Thus, an elliptic point is one where $K > 0$. In other words, normal sections through the point are all convex or all concave, the surface does not intersect its tangent plane at this point. An example is any point of an ellipsoid. A parabolic point is one where $K = 0$. An example is any point of a cylinder. A hyperbolic point (or saddle point) is one where $K < 0$. In other words, there are both convex and concave normal sections, the surface intersects its tangent plane. An example is any point of a hyperboloid of one sheet.

3.3 Example

Let the surface be represented by the vector $\mathbf{r}=\{x,y,z(x,y)\}$. Then the derivatives are

$$\mathbf{r}_1 = \{1,0,z_x\}$$

$$\mathbf{r}_2 = \{0,1,z_y\}$$

$$\mathbf{r}_1 \times \mathbf{r}_2 = \{-z_x,-z_y,1\}$$

$$|\mathbf{r}_1 \times \mathbf{r}_2| = (1+z_x^2+z_y^2)^{1/2}.$$

Since $\mathbf{r}_1$ and $\mathbf{r}_2$ are tangents to the parametric curves, the normal to the surface lies

perpendicular to both of them. Thus the unit normal to the surface is

$$\mathbf{n} = (1+z_x^2+z_y^2)^{-1/2} \{-z_x,-z_y,1\}.$$

Hence

$$E = 1+z_x^2$$
$$F = z_x z_y$$
$$G = 1+z_y^2$$
$$H^2 = 1+z_x^2+z_y^2$$

Similarly

$$\mathbf{r}_{11} = \{0,0,z_{xx}\}$$
$$\mathbf{r}_{12} = \{0,0,z_{xy}\}$$
$$\mathbf{r}_{22} = \{0,0,z_{yy}\}$$
$$L = z_{xx}/H$$
$$M = z_{xy}/H$$
$$N = z_{yy}/H$$
$$T^2 = H^{-2} (z_{xx}z_{yy}-z_{xy}^2)$$

Thus, for calculation purposes,

$$J = \partial/\partial x \, (z_x/(1+z_x^2+z_y^2)^{1/2}) + \partial/\partial y \, (z_y/(1+z_x^2+z_y^2)^{1/2})$$
$$K = (1+z_x^2+z_y^2)^{-2} \, (z_{xx}z_{yy}-z_{xy}^2) \, .$$

This example suggests that there may be another useful representation of J, to which we now turn.

The divergence of a vector is defined (as in Weatherburn) as

$$\mathrm{div}\mathbf{F} = H^{-2}\mathbf{r}_1\cdot(GF_u-FF_v) + H^{-2}\mathbf{r}_2\cdot(EF_v-FF_u) \, .$$

This can be used to show that

$$\mathrm{div}\, \mathbf{n} = -J \, .$$

In the case of a surface with parameters x, y

$$\mathrm{div}\, \mathbf{n} = \nabla\cdot\mathbf{n} = \partial n_1/\partial x + \partial n_2/\partial y \qquad \mathbf{n}=\{n_1,n_2,n_3\} \, .$$

As well, if the Laplacian is denoted by ∇^2, then

$$2K = \mathbf{n}\cdot\nabla^2\mathbf{n} + (\nabla\cdot\mathbf{n})^2 \, .$$

For the case of $\mathbf{r} = \{x,y,z(x,y)\}$

$$\nabla^2\mathbf{n} = \{\nabla^2 n_1, \nabla^2 n_2, \nabla^2 n_3\}$$
$$\nabla^2 n_i = \partial^2 n_i/\partial x^2 + \partial^2 n_i/\partial y^2 \, .$$

We can now make more precise the earlier suggestion that first curvature of a surface is analogous to the curvature of a curve, and second curvature is analogous to the torsion. For a surface

$$\mathrm{div}\, \mathbf{n} = \nabla\cdot\mathbf{n} = -J$$

$$\nabla^2 r = Jn$$
$$n \cdot \nabla^2 n + (\nabla \cdot n)^2 = 2K$$

For a curve, one has the analogous equations, using a one parameter version of the del operator

$$\nabla = t d\cdot/ds .$$

Thus

$$\text{div } n = \nabla \cdot n = -\kappa$$
$$\nabla^2 r = \kappa n$$
$$n \cdot \nabla^2 n + (\nabla \cdot n)^2 = -\tau^2 .$$

Thus J is analogous to κ and 2K is analogous to $-\tau^2$.

3.4 Extremal Conditions

One of the initial constraints applied to the problem of contructing the surface was based on the manner in which the zero-crossing contours of the stereo program were formed. In essence, it claimed that for portions of the image not associated with a zero-crossing, the underlying surface must change in a "reasonable" manner. What does this imply about the surface itself?

Our assumption requires that between zero-crossing contours, the surface should change as little as possible, and in particular should not have any inflection points in depth that are not necessary. This is since such inflections should give rise to other zero-crossings which are not indicated in the stereo output. At the same time, the boundary values along the zero-crossing contours, around some portion of the surface, will impose a certain amount of intrinsic curvature to any surface fitting them. One method for finding a surface which will fit the boundary conditions along the zero-crossing contours, and yet will not curve excessively, is to require that the average curvature at any point on the surface be minimal.

A theorem due to Euler states that if κ_n is the curvature of the curve obtained by taking a normal section of the surface with a plane whose orientation relative to one of the principal directions is β, then

$$\kappa_n = \kappa_a \cos^2\beta + \kappa_b \sin^2\beta .$$

By taking all possible normal sections and integrating, one finds that the average curvature at a point is given by J/2. Hence, one possibile method for obtaining a surface which fits a set of boundary conditions and has a particular extremal property related to the curvature of the surface is:

1. Find the surface fitting the boundary conditions which minimizes

$$\iint J^2 dx\ dy = \iint (\nabla \cdot n)^2 dx\ dy = \iint (\kappa_a + \kappa_b)^2 dx\ dy.$$

A surface which minimizes the above integral for some set of boundary conditions is a surface which minimizes the average curvature of the surface at every point.

How well does this condition satisfy our constraints? As we shall see, it is possible to construct a computational scheme which computes a surface satisfying this condition and yet is consistent with the notion of a local parallel algorithm. As to whether the surfaces so constructed meet our condition of "well-behaved" surfaces we note the following. In the case of boundary conditions which are consistent with a simple surface, such as a plane, a cylinder, a sphere or even an ellipsoid, the above condition will lead to the correct surface in each case. This certainly meets our requirement that the surface behave smoothly and not change any more than required. However, if the surface determined by the boundary conditions and the above requirement has a hyperbolic point, it is possible to minimize J^2 at such a point without minimizing the principle curvatures. This is undesirable since it could lead to surfaces which violate our requirement of no extremes except at zero-crossings.

One manner of altering the above condition to handle hyperbolic points as well is the following:

2. Find the surface fitting the boundary conditions which minimizes

$$\iint {\kappa_a}^2 + {\kappa_b}^2 dx\ dy = -\iint (n \cdot \nabla^2 n) dx\ dy.$$

Here, it is no longer possible to minimize the integrand without minimizing the principle curvatures as well. Thus we seem to meet our analytic constraint better than the previous suggestion. However, finding a method of computing the surface is not as easy. Note that

$${\kappa_a}^2 + {\kappa_b}^2 = (\kappa_a + \kappa_b)^2 - 2\kappa_a\kappa_b = J^2 - 2K = -n \cdot \nabla^2 n\ .$$

This gives an alternate form of the integrand to be minimized. However, the conditions which will minimize such an integral give rise to a non-linear system of equations. Whereas it is possible to devise a computational scheme which is consistent with the notion of a local parallel algorithm, it is difficult to prove that such an algorithm will converge to the correct surface. Such an algorithm may be undesirable for this reason.

We have required that the surface which fits the boundary conditions curve as little as possible. We achieved this in the first case by requiring that the average curvature at every point be as small as possible. An alternate method of keeping the curvature of the surface minimal is as follows.

Consider a small segment of the surface, surrounding a point P. For each point within the segment, translate the unit normal at that point to the origin of the coordinate system. This results in the inscription of some region(s) of the unit sphere. If the surface is smooth in the region of P, then the inscribed portion of the unit sphere is a single connected region. If one considers the ratio of the area of the inscribed region to the area of the original region, and takes the limit as the region surrounding P tends to the point itself, one obtains the Gaussian curvature at P. Thus the Gaussian curvature also measures the amount of "bending" of the surface at a point. Thus a possible constraint is:

3. Find the surface fitting the boundary condition which minimizes

$$\iint K^2 dx\, dy = \iint \kappa_o^2 \kappa_b^2 dx\, dy .$$

Minimizing the above integral results in a surface which reduces the "bending", as measured by the Gaussian curvature, to a minimum.

There may be other extremal conditions that can be applied. The critical point is that they ensure that the surface which satisfies them is consistent with the constraints previously derived. That is, they must not have additional inflections other than at zero-crossings, they must spread the curvature of the surface in a smooth manner over the surface, and they must fit the stereo boundary conditions along the zero-crossings contours.

Most of the extremal conditions listed above are phrased in terms of the surface normal at points on the surface. Hence they are best suited for constructing a surface when the boundary conditions are specified in terms of surface orientation. However, one can also contruct a surface when the boundary conditions are specified in terms of distance. The next section examines this case.

4. Surface Patches

We have indicated that our basic intention is to fit a smooth surface to a set of boundary conditions. Since the number of possible surfaces is infinite and varied, we have restricted our attention to those surfaces which satisfy certain extremal conditions with respect to surface curvature. This section outlines a particular method for fitting smooth surfaces to boundary conditions. This method is especially suited to the task of fitting a surface to boundary conditions involving depth information.

Coons [1967] has developed a method for piecing together patches of a smooth surface in such a way as to ensure continuity of both the function describing the surface and its derivatives up to some order r. The method is described below.

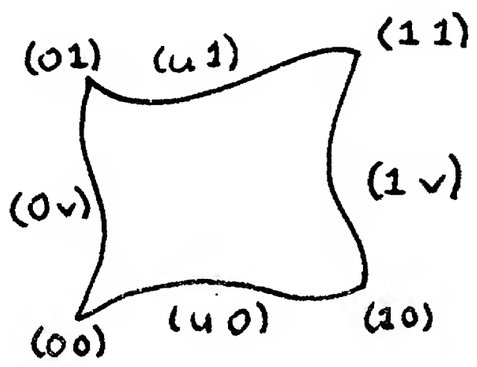
Suppose that the surface is parameterized in terms of u,v so that in a Cartesian coordinate system the coordinates are given by

$$x = f(u,v)$$
$$y = g(u,v)$$
$$z = h(u,v) .$$

Utilizing Coons' notation, let us denote any point on the surface for the three dimensional case by

$$(uv) = [f(u,v), g(u,v), h(u,v)]$$

Furthermore we can scale u and v so that they range from 0 to 1. So a surface patch is a segment bounded by four space curves, namely (0v), (1v), (u0), (u1).



In order to define a smooth surface segment (or patch) between these boundary curves, we will blend the boundary curves together in a smooth manner. Let F_0 and F_1 be two blending functions. Then the surface equation is given by

$$(uv) = \sum_i (iv)F_i(u) + \sum_j (u,j)F_j(v) - \sum_i\sum_j (ij)F_i(u)F_j(v) .$$

The F's are restricted such that

$$F_0(0) = 1 \qquad F_0(1) = 0$$
$$F_1(0) = 0 \qquad F_1(1) = 1 .$$

This ensures that the surface defined by the surface equation above contains its boundaries and corner points. In general, the functions F_0 and F_1 are taken to be continuous and monotonic, but this is not critical. The boundary curves should be closed and continuous.

By imposing the additional restrictions

$$F_0'(0) = 0 \qquad F_0'(1) = 0$$
$$F_1'(0) = 0 \qquad F_1'(1) = 0$$

the slope of the surface across a boundary has the form

$$(u0)_v = (00)_v F_0(u) + (10)_v F_1(u) .$$

Thus the slope across the boundaries depends on the two end tangent vectors across the boundary and on the blending functions. It is independent of the actual shape of the boundary curve. Thus, any two patches which share a common boundary will be continuous in slope across this common boundary under the above restrictions on the blending functions. Similarly, if the second order derivatives satisfy

$$F_0''(0) = 0 \qquad F_0''(1) = 0$$
$$F_1''(0) = 0 \qquad F_1''(1) = 0$$

then the patches will match in the second derivative across common boundaries.

Differentiation indicates that

$$(00)_{uv} = (01)_{uv} = (10)_{uv} = (00)_{uv} = 0 .$$

In other words, the cross-derivative or twist vectors at the patch corners are all zero.

Not all the surfaces with which we must deal will have cross-boundary slopes of the above form, nor will they all have zero twist vectors at the corners of the patches. In such cases, a correction surface should be added to the original surface to account for this. The function of this correction surface is to correct the slopes across the patch boundary and the corner twist vectors without changing the shape of the boundary curves. In other words, the correction surface changes only slope and higher order conditions. This correction surface is defined by

$$(uv) = \sum_i (iv)_u G_i(u) + \sum_j (uj)_v G_j(v) - \sum_i \sum_j (ij)_{uv} G_i(u) G_j(v) .$$

Adding the two surfaces together, where we denote the initial surface by f(u,v) and the correction surface by g(u,v), we see that the slope across the patch is given by

$$(u0)_v = f(u,0)_v + g(u,0)_v .$$

Hence, we choose $g(u,0)_v$ and the other cross-boundary slope functions in g(u,v) such that

$$g(u,0)_v = (u0)_v - (00)_v F_0(u) - (10)_v F_1(u) .$$

The functions G_0 and G_1 are the slope blending functions and have the following end conditions in order to ensure that the correction surface does not alter the patch boundaries but has the required cross-boundary slopes.

$$G_0(0) = 0 \quad G_0(1) = 0 \quad G_0'(0) = 1 \quad G_0'(1) = 0$$
$$G_1(0) = 0 \quad G_1(1) = 0 \quad G_1'(0) = 0 \quad G_1'(1) = 1 \ .$$

In a similar manner, one can apply higher derivative corrections.

The entire surface equation can be formulated in terms of tensors

$$(uv) = -[-1\ F_0(u)\ F_1(u)\ G_0(u)\ G_1(u)] \begin{bmatrix} 0 & (u0) & (u1) & (u0)_v & (u1)_v \\ (0v) & (00) & (01) & (00)_v & (01)_v \\ (1v) & (10) & (11) & (10)_v & (11)_v \\ (0v)_u & (00)_u & (01)_u & (00)_{uv} & (01)_{uv} \\ (1v)_u & (10)_u & (11)_u & (10)_{uv} & (11)_{uv} \end{bmatrix} \begin{bmatrix} -1 \\ F_0(v) \\ F_1(v) \\ G_0(v) \\ G_1(v) \end{bmatrix}$$

This is the equation for a slope-matching, slope-continuous surface patch with arbitrary boundaries and arbitrary slope across the boundaries.

Note that the correction surface appears to be essential in order to achieve a "fair" interpolated surface. This is because the original surface equation had zero twist vectors, yet most doubly curved surfaces do not. Forrest observes [1968]:

> "It is said that where a series of Coons first canonical form patches [i.e. only the basic surface equation] are fitted to an array of points on a car body panel, and the panel is cut by numerical control, the patches can be distinguished on the panel by a series of flattenings or local distortions; the overall surface is smooth but clearly not fair."

We have not yet specified the form of the blending functions, and to this we now turn. Let

$$[u_1\ u_2\ u_3\ u_4]$$

be a vector whose elements are a set of linearly independent functions of u. Then the blending functions can be specified by

$$[F_0(u)\ F_1(u)\ G_0(u)\ G_1(u)] = [u_1\ u_2\ u_3\ u_4]\ M \ .$$

Thus we want to specify M such that

$$M^{-1} = \begin{bmatrix} u_1|_{u=0} & u_2|_{u=0} & u_3|_{u=0} & u_4|_{u=0} \\ u_1|_{u=1} & u_2|_{u=1} & u_3|_{u=1} & u_4|_{u=1} \\ du_1/du|_{u=0} & du_2/du|_{u=0} & du_3/du|_{u=0} & du_4/du|_{u=0} \\ du_1/du|_{u=1} & du_2/du|_{u=1} & du_3/du|_{u=1} & du_4/du|_{u=1} \end{bmatrix}$$

For example, we could use a cubic basis vector, where

$$[u_1\ u_2\ u_3\ u_4] = [u^3\ u^2\ u\ 1]\ .$$

In this case,

$$M^{-1} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 3 & 2 & 1 & 0 \end{bmatrix} \qquad M = \begin{bmatrix} 2 & -2 & 1 & 1 \\ -3 & 3 & -2 & -1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

Then we have

$$F_0(u) = 2u^3 - 3u^2 + 1$$
$$F_1(u) = -2u^3 + 3u^2$$
$$G_0(u) = u^3 - 2u^2 + u$$
$$G_1(u) = u^3 - u^2\ .$$

If both basis vectors, $[u_1\ u_2\ u_3\ u_4]$ and $[v_1\ v_2\ v_3\ v_4]$, are taken to be cubic basis vectors, then the resultant patch will be a bicubic surface, which is the two-dimensional analog of the cubic spline. Of course, many other blending functions are possible, each one resulting in different types of smooth surface patches.

5. Convergence Methods

The previous two sections have dealt with the problem of how to reconstruct a surface having particular properties from a set of boundary conditions. The reconstruction itself has not been formulated, and we now examine methods for computing the actual value of the surface at various points, subject to the reconstruction methods, of course. Our point of view in handling this task will be to compute the value of the surface with as local a support as possible.

Most of the extremal constraints on the curvature of a surface can be turned into a set of linear difference equations, based on a two-dimensional grid. Similarly, the Coons surface patch equation can be turned into a set of linear difference equations based on a two-dimensional grid. Thus, we examine iterative methods for solving a set of linear equations. We look at iterative methods rather than elimination methods because we seek local, rather than global, computations.

Given a system of linear equations

$$Ax = c$$

which is nonsingular, we want to find a sequence x_k such that x_k converges to $A^{-1}c$. An iteration is said to be of degree r if x_k is a function of A, c, x_{k-1}, ..., x_{k-r}. Usually, r=1 so that

$$x_k=F_k(A, c, x_{k-1})$$

If F_k is independent of k then this is a stationary iteration. If F_k is a linear function then the iteration is linear.

5.1 Linear Case

Suppose that

$$F_k(A,c,x_{k-1}) = H_k x_{k-1} + v_k$$

where H_k is a function of A and c. The solution should be invariant so

$$A^{-1}c=H_kA^{-1}c+v_k$$

which implies

$$v_k = M_kc \quad \text{where} \quad M_k = (I-H_k)A^{-1} .$$

Thus,

$$x_k = H_kx_{k-1} + M_kc \quad \text{and} \quad H_k + M_kA = I .$$

The error associated with each iteration is

$$e_k = x_k-A^{-1}c \quad \text{where} \quad e_k = H_ke_{k-1}.$$

Then the system given above is convergent for a given initial error e_0 if and only if K_kx converges to 0 for any x, where $K_k=H_kH_{k-1}...H_1$. For stationary linear iterations,

$H_k=H$ and $K_k=H^k$. Thus we need to check that H^kx converges to 0. It can be shown that H^kx converges to 0 for arbitrary x if and only if each eigenvalue λ_i of H is such that $|\lambda_i|<1$.

Bearing this in mind, we can now develop particular methods of iteration.

5.2 Simultaneous Displacements (Jacobi, 1845)

Given $Ax=c$ let $A=E+D+F$ where D is a diagonal matrix, E is a lower triangular matrix, F is an upper triangular matrix and $a_{ii}\neq 0$ for all i. Suppose we are given a trial solution x_0 and we have found x_{k-1}, $(k = 1,2,\ldots)$, then we solve

$$Ex_{k-1} + Dx_k + Fx_{k-1} = c$$
$$\text{or} \quad x_k = Hx_{k-1} + Mc$$

where $H = -D^{-1}(E+F)$ and $M = D^{-1}$. In order for this iterative process to converge, it is necessary and sufficient that the eigenvalues λ_i of H be bounded in modulus by 1. Note that if $D = \delta I$ then an eigenvalue λ_i of H corresponds to an eigenvalue $(1-\lambda_i)\delta$ of A.

One set of sufficient conditions is given by the following result, due to Collatz [1950].

If A has diagonal dominance and is not reducible, then the method of simultaneous displacements converges. Note that A is diagonally dominant if and only if

$$\Sigma_{i\neq j}|a_{ij}| \leq a_{jj} \quad \text{for all } j$$

with strict inequality for at least one j. A is reducible if the set $\{1,\ldots N\}$ is the union of two nonempty sets S and T such that $a_{ij}=0$ for all $i\in S$, $j\in T$. In other words, there is a permutation matrix Π such that

$$\Pi A\Pi^T = \begin{bmatrix} A_1 & 0 \\ A_2 & A_3 \end{bmatrix}.$$

5.3 Gradient method

If A is symmetric and definite, we can use the gradient method. Let

$$E(x) = x^TAx - 2c^Tx.$$

Since $E(x) = (x-A^{-1}c)^TA(x-A^{-1}c) - c^TA^{-1}c$ it follows that $E(x)$ attains a minimum value $-c^TA^{-1}c$ precisely when $x=A^{-1}c$. Solving the system $Ax=c$ is then equivalent to finding the x which minimizes $E(x)$.

Given $x_0,\ldots,x_{k-1}$, compute the gradient direction of $E(x)$ at x_{k-1}

$$-\nabla E(x)|_{x=x_{k-1}} = -2(Ax_{k-1}-c) = 2r_{k-1}$$

where r_k is the residual. Since $E(x)$ decreases in the direction of r_{k-1}, we choose

$$x_k = x_{k-1}+\alpha_{k-1}r_{k-1} .$$

We now need to select α_{k-1}. One possibility is to choose the optimal value, which can be shown to be

$$\frac{r_{k-1}^T r_{k-1}}{r_{k-1}^T A r_{k-1}}$$

A second possibility is to let $\alpha_k=\alpha$ for all k. Then the system becomes

$$x_k = x_{k-1} + \alpha(c-Ax_{k-1})$$

and the error term is $e_k=x_k-A^{-1}c$ such that

$$e_k = e_{k-1} + \alpha(-Ae_{k-1}) = (I-\alpha A)e_{k-1} .$$

So we have a stationary linear iteration with $H=I-\alpha A$. Let μ_i be the eigenvalues of A, λ_i those of H. Then $\lambda_i=1-\alpha\mu_i$. Since A is positive definite, we know that $\mu_i > 0$ for all i. The requirement of $|\lambda_i| < 1$ for convergence then becomes

$$0 < \alpha < 2/\max_i \mu_i .$$

For the fastest convergence, we must minimize $\max_i |\lambda_i|$, so we choose α such that $\max_i |1-\alpha\mu_i|$ is a minimum.

If we know that the μ_i lie in the interval [a,b], $0<a<b<\infty$, then $|1-\alpha\mu_i|$ is a maximum at the endpoints. So the best choice is $1-\alpha a = -(1-\alpha b)$ which implies $1/\alpha = (a+b)/2$. In this case,

$$|1-\alpha\mu_i| \leq (b-a)/(b+a) < 1 .$$

The gradient method is identical to the method of simultaneous displacements whenever A is symmetric and definite with a scalar matrix as diagonal ($D=\delta I$).

5.4 Richardson's method

We have seen that in the gradient method

$$x_k = x_{k-1} + \alpha_{k-1}r_{k-1} .$$

This suggests that the relaxation parameter α may be a function of k.

Suppose A is positive definite and, as before, the error term is

$$e_k = (I-\alpha_{k-1}A)e_{k-1}$$

$$e_k = (I-\alpha_0)(I-\alpha_1 A)\ldots(I-\alpha_{k-1}A)e_0 .$$

Thus we have

$$e_k = P_k(A)e_0$$

where

$$P_k(x) = \Pi_{0\le i\le k-1} (1-\alpha_i x)$$

is a polynomial in x such that $P_k(0) = 1$. Furthermore, the zeros of P_k are $1/\alpha_i$. So choosing the constants α_i is the same as choosing a polynomial P_k of degree k such that $P_k(0)=1$. There are a number of ways of doing so; the following is one of the more useful.

If the eigenvalues of A are μ_i then those of $P_k(A)$ are $P_k(\mu_i)$. Suppose that $\mu_i \epsilon [a,b]$ where $0<a<b<\infty$. Let

$$e_0 = \Sigma_{1\le i\le N} \gamma_i x_i$$

where the x_i's are the eigenvectors of A, and form a basis. Then

$$e_k = \Sigma_{1\le i\le N} \gamma_i P_k(\mu_i) x_i \ .$$

This suggests that one method for ensuring that the error terms vanish quickly is to make $P_k(\mu_i)$ small. It can be shown [Forsythe 1960, p227] that one can minimize $\max_{a\le x\le b} |P_k(x)|$ by choosing

$$P_k(x) = T_k[(b+a-2x)/(b-a)]/T_k[(b+a)/(b-a)]$$

where $T_k(y)$ is the Chebyshev polynomial and is given by

$$T_k(y) = \cos[k \arccos y)] \ .$$

Thus $P_k(x)$ is simply the Chebyshev polynomial adjusted to the interval [a,b] and scaled to satisfy $P_k(0)=1$. Let $y_0 = (b+a)/(b-a)$. Then

$$2T_k(y) = [y + y^2-1)^{1/2}]^k + [y - (y^2-1)^{1/2}]^k \ .$$

Thus, as k tends to infinity,

$$\max_{a\le x\le b} |P_k(x)| \le 2(y_0 - (y_0^2-1)^{1/2})^k$$

The average rate of convergence is then bounded by

$$- \log(\|e_k\|/\|e_0\|)^{1/k} \le -1/k \log 2 + \log (y_0 + (y_0^2 - 1)^{1/2} \ .$$

To compare this method to the simultaneous displacements method, let $P = \max \mu_i / \min \mu_i$, where μ_i are the eigenvalues of A, which is positive definite. Then the rate of convergence for the Richardson method is

$$\log [y_0 + (y_0^2-1)^{1/2}] = 2(a/b)^{1/2} \le 2P^{-1/2} \ .$$

The rate of convergence for the method of simultaneous displacements is

$$\log y_0 \simeq 2a/b \le 2P^{-1} \ .$$

For a process of degree 1, we cannot in general let k tend to infinity. If we use a process such as

$$x_k = x_{k-1} + \alpha_{k-1} r_{k-1}$$

we need to know the value of α_i. Since the α_i are the reciprocals of the zeros of $P_k(x)$, we can fix k at say K = 20, then determine α_i from the roots of the Chebyshev

polynomial and compute $x_1,...,x_k$. If the error term e_K is not small enough, one can use the same values of α_i to make another cycle of K steps. Such a process is called semi-iterative.

5.5 Successive Displacements (Gauss-Seidel [1874])

The method of successive displacements takes advantage of the computation of the components of the vector as soon as they become available. Thus, the process is

$$Ex_k + Dx_k + Fx_{k-1} = c$$
$$x_k = -(D+E)^{-1}Fx_{k-1} + (D+E)^{-1}c .$$

As with the case of simultaneous displacements, the following results hold.

(Collatz [1950]) If A has diagonal dominance and is not reducible then the method of successive displacements converges.

(Reich [1949]) If A is symmetric and nonsingular and $a_{ii}>0$, then the method of successive displacements converges for all initial states x_0 if and only if A is positive definite.

5.6 Overrelaxation

One can take the method of successive displacements a step further. For processes such that a_{ii} is nonzero for all i, consider

$$(E+\omega^{-1}D)x_k + [F+(1-\omega^{-1})D]x_{k-1} = c .$$

The parameter ω is a relaxation factor. If $0<\omega<1$, then the system is underrelaxed. If $1<\omega<2$ then the system is overrelaxed. In particular, if A is symmetric, the diagonal elements are positive and the off-diagonal elements are non-positive, then the following is true.

If the method of succesive displacements converges then the method of successive overrelaxation or underrelaxation converges for all ω such that $0<\omega<2$. If the method of successive relaxation converges for any $0<\omega<2$, then the method of simultaneous displacements converges.

The value to be assigned to the relaxation factor ω can be determined as follows. Rescale the matrix A such that the diagonal elements are 1. Let A = -E+I-F. Let μ_i be the eigenvalues of A and let λ_i be the eigenvalues of the method of

simultaneous displacements associated with A. Then $\lambda_i = 1 - \mu_i$. Let $\nu_1^{(w)}$ be the dominant eigenvalue of the successive overrelaxation method. Then if

$$\omega_b = 2/[1 + (1-\lambda_1^2)^{1/2}]$$

the following relationship holds (Kahan [1958]).

$$\omega_b - 1 \leq |\nu_1^{(w)}| < (\omega_b - 1)^{1/2} .$$

Although the optimal relaxation factor may not be known, ω_b may be close to it, especially when $\omega_b \simeq 2$.

Although any of the above methods may be used to solve a system of linear equations, there are various tradeoffs associated with each of them. For example, although the method of simultaneous displacements tends to have slow convergence, it can often be implemented by a network of processors which are very local in support. On the other hand, the same system of equations could be solved by the method of successive displacements. Here, the convergence is much faster, but the support of the individual processors is much more global.

6. Reconstruction of Surfaces

Previously, a number of possible extremal conditions were suggested for the construction of a surface. As well, various iterative convergence methods were suggested for actually computing the surface. We now consider the effects of combining these two factors.

Consider the case of finding the surface which fits the boundary conditions and minimizes

$$\iint J^2 \, dx \, dy = \iint (\kappa_a + \kappa_b)^2 \, dx \, dy \, .$$

Note that this criterion is well suited for the case of interpolating a surface from surface normal information, since any constraints derived from such a condition directly relate to the value of the surface normal on the surface.

Applying the calculus of variations to this integral equation results in the necessary conditions that the partial derivatives vanish.

$$\partial/\partial x \, [\partial n_1/\partial x + \partial n_2/\partial y] = \partial/\partial y \, [\partial n_1/\partial x + \partial n_2/\partial y] = 0 \, .$$

These Euler conditions derived from the calculus of variations can be transformed into a set of discrete conditions and thus into a set of linear difference equations. This allows one to use the convergence methods discussed above to reconstruct the surface.

Since the partial derivatives above are bivariate, the iterative matrix of the set of linear difference equations is block tridiagonal. As a consequence, the matrix as it stands can be shown to be divergent. However, if we transform the system into a one-dimensional system by means of some simple matrix manipulation, the system becomes convergent, and moreover, is very local in nature. Thus, the system for minimizing the integral of J^2 satisfies our condition of having computations which are local in support. Moreover, the resulting process could easily be implemented in a parallel network of local support processors.

Any surface constructed under such a scheme can be shown to be locally composed either of planes, cylinders, spheres or ellipsoids. Thus the only basic type of surface not handled by this method is hyperboloids.

It was suggested earlier that one method for overcoming this handicap would be to minimize the integral of $\kappa_a^2 + \kappa_b^2$ since such a surface could contain hyperbolic points. When one applies the calculus of variations to such a system, a nonlinear set of equations are obtained, and the convergence methods described above are no longer applicable.

Thus, given the constraint of requiring a computational system which is local in support and can be computed by iterative methods, a good candidate for reconstructing surfaces which fit a certain set of boundary conditions and elsewhere contain no inflection points is to find the surface minimizing

$$\int\int J^2 dx\ dy = \int\int (\nabla \cdot n)^2 dx\ dy = \int\int (\kappa_a + \kappa_b)^2 dx\ dy.$$

As a second example, let us consider how to fit a smooth surface to a set of boundary conditions, using the Coons surface patch method. For simplicity, the boundary is taken to be a square along which the distance to the surface is known up to a scale factor. The interior of the square is of size n. The basis vectors for the Coons patch are cubics and the underlying parameters are taken to be the axis variables, x and y.

We treat the underlying system as a two-dimensional grid so that we must compute the depth of the surface at each point on an n*n grid. Suppose we know the values of the surface at the corners of a particular patch, and we wish to compute the value for the center of the patch, that is for $x = y = 1/2$. Then $F_0(1/2) = F_1(1/2) = 1/2$ and $G_0(1/2) = -G_1(1/2) = 1/8$.

The derivatives can be approximated by discrete differences between points on the grid. By evaluating the tensor form of the surface equation, we find that the value of the midpoint is given by

$$(.5\ .5) = .5\{(0\ .5) + (1\ .5) + (.5\ 0) + (.5\ 1)\}$$
$$-.25\{(0\ 0) + (0\ 1) + (1\ 0) + (1\ 1)\}\ .$$

We can treat each surface patch as that patch defined by a 3*3 piece of the grid. Then, combining the surface equation for each point of the grid, we obtain the linear system

$$Ax = c$$

where the constant vector c is obtained by applying the above equation to the boundary points for the case of a point next to the boundary. Here the matrix A has a tridiagonal block form

$$A = \begin{bmatrix} A_1 & -.5A_1 & & & \\ -.5A_1 & A_1 & -.5A_1 & & \\ & & \cdots & & \\ & & & -.5A_1 & A_1 \end{bmatrix}.$$

$$A_1 = \begin{bmatrix} 1 & -.5 & & \\ -.5 & 1 & -.5 & \\ & \cdots & & \\ & & -.5 & 1 \end{bmatrix}$$

Applying the Jacobi method of simultaneous displacements to this block system, yields the system

$$x_{k+1} = Bx_k + D^{-1}c$$

with

$$B = .5\begin{bmatrix} 0 & I & & \\ I & 0 & I & \\ & \cdots & & \\ & & I & 0 \end{bmatrix}$$

$$D^{-1} = \begin{bmatrix} A_1^{-1} & & & \\ & A_1^{-1} & & \\ & & \cdots & \\ & & & A_1^{-1} \end{bmatrix}$$

In other words, at each stage of the iteration, each point in the grid assumes the average of its neighbours in the y direction, plus some constant term which reflects the influence of the boundaries in the x direction. The constant term is given by $D^{-1}c$ where $A_1^{-1} = [a_{ij}]$ with

$$a_{ij} = 2/(n+1)\ i\ (n+1-j) \quad i \leq j$$
$$= 2/(n+1)\ j\ (n+1-i) \quad i \geq j$$

Provided the boundary of the entire piece of the surface is closed, i.e. fixed depth values are known along a closed contour surrounding the piece of the surface, the matrix B satisfies the Chebyshev recurrence relation and hence has eigenvalues $\cos(k\pi/n+1)$ for $k = 1,2,\ldots,n$. Thus the convergence rate for this process is $\cos(\pi/n+1)$, and the asymptotic rate is roughly n^{-2}, which is slow. However, note that the support of each computation is small, since each point of the grid obtains its value by examining the values of only two neighbours.

7. Summary

This paper has been a discussion of several pieces of mathematics relevant to the interpolation of a surface from a set of boundary conditions. The first section outlined the problem as posed by the Marr & Poggio theory of stereopsis. It was shown that the stereo process imposes both explicit constraints along the zero-crossing contours obtained by processing the image, and implicit constraints elsewhere. These implicit constraints in particular included the requirement that the surface not change its curvature in a radical manner for locations not associated with zero-crossings. The second section dealt with some relevant details of differential geoemetry, in particular, those aspects dealing with the curvature of surfaces. The third section outlined the method of Coons for fitting fair surfaces to boundary conditions. The fourth section sketched various iterative schemes for computing the surface, and the fifth section tied all of these notions together and sketched the reconstruction method.

8. References

Collatz, L.[1950]: Uber die Konvergenzkriterien bei Iterationsverfahren fur lineare Gleichungssysteme, Math. Z., vol 53, pp149-161.

Coons, S.A.[1967]: Surfaces for Computer-aided Design of Space Forms, MAC-TR-41, Project MAC, Massachusetts Institute of Technology.

Forrest, A.R.[1968]: Curves and Surfaces for Computer-Aided Design, Joint Computer-Aided Design Group, University of Cambridge.

Forsythe, G. and W. Wasow[1960]: Finite-difference Methods for Partial Differential Equations, John Wiley & Sons, New York.

Horn, B.K.P.[1970]: Shape form Shading: A Method for Obtaining the Shape of a Smooth Object from One View, MAC-TR-70, Project MAC, Massachusetts Institute of Technology.

Horn, B.K.P.[1975]: Obtaining Shape form Shading Information, in The Psychology of Computer Vision, P.H. Winston (ed.), McGraw-Hill, pp115-155.

Jacobi, C.G.J.[1845]: Ueber eine neue Auflosungsart der bei der Methode der kleinsten Quadrate vorkommenden linearen Gleichungen, Astr. Nachr., vol 22, No. 523, pp297-306.

Kahan, W.[1958]: Gauss-Seidel Methods of Solving Large Systems of Linear Equations, Thesis, Toronto.

Marr, D.[1974]: A Note on the Computation of Binocular Disparity in a Symbolic, Low-level Visual Processor, MIT AI Memo 327, Massachusetts Institute of Technology.

Marr, D.[1976]: Early Processing of Visual Information Phil. Trans. R. Soc. B., vol. 226, pp483-524.

Marr, D. and E. Hildreth[1979]: Theory of Edge Detection, submitted to Proc. R. Soc. Also available as MIT AI Memo 518, Massachusetts Institute of Technology.

Marr, D. and T. Poggio[1977]: A Theory of Human Stereo Vision, submitted to Proc. R. Soc. Also available as MIT AI Memo 451, Massachusetts Institute of Technology.

Reich, E.[1949]: On the Convergence of the Classical Iterative Method of Solving Linear Simultaneous Equations, Ann. Math. statis., vol. 20, pp448-451.

Seidel, L.[1874]: Ueber ein Verfahren, die Gleichungen, auf welche die Methode der kleinsten Quadrate fuhrt, sowie lineare Gleichungen uberhaupt, durch successive Annaherung aufzulosen, Abh. Math.-Phys. Kl., Bayerische Akad. Wiss Munchen, vol 11 (III), pp81-108.

Weatherburn, C.E.[1927]: Differential Geometry of Three Dimensions, Cambridge Press.

Wilson, H.R. and J.R. Bergen[1979]: A Four Mechanism Model for Spatial Vision (In the press).

Wilson, H.R. and S.C. Giese[1977]: Threshold Visibility of Frequency Gradient Patterns, Vision Res. 17, 1177-1190

Woodham, R.J.[1978]: Reflectance Map Techniques for Analyzing Surface Defects in Metal Castings, MIT AI Technical Report 457, Massachusetts Institute of Technology.